MBE Advance Access published April 10, 2015

Type: Article

Section: Discoveries

Single geographic origin of a widespread autotetraploid *Arabidopsis arenosa* lineage followed by interploidy admixture

Brian Arnold[1], Sang-Tae Kim[2,3], Kirsten Bomblies[1]

1. Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA
2. Max Planck Institute for Developmental Biology, Tübingen, Germany
3. Current address: Center for Genome Engineering, Institute for Basic Science, Daejeon, South Korea.

ABSTRACT

Whole-genome duplication (WGD), which leads to polyploidy, has been implicated in speciation and biological novelty. In plants, many species exhibit ploidy variation, which is likely representative of an early stage in the evolution of polyploid lineages. To understand the evolution of such multiploidy systems, we must address questions such as whether polyploid lineage(s) had a single or multiple origins, whether admixture occurs between ploidies, and the timescale over which ploidy variation affects the evolution of populations. Here we analyze three genomic datasets using nonparametric and parametric analyses, including coalescent-based methods, to study the evolutionary history of a geographically widespread autotetraploid variant of *Arabidopsis arenosa*, a new model system for understanding the molecular basis of autopolyploid evolution. Autotetraploid *A. arenosa* populations are widely distributed across much of Northern and Central Europe, while diploids occur in Eastern Europe and along the southern Baltic coast; the two ploidies overlap in the Carpathian Mountains. We find that the widespread autotetraploid populations we sampled likely arose from a single ancestral population ~11,000-30,000 generations ago in the Northern Carpathians, where its closest extant diploid relatives are found today. Afterward, the tetraploid population split into at least four major lineages that colonized much of Europe. Reconstructions of population history suggest substantial interploidy admixture occurred in both directions, but only among geographically proximal populations. We find two cases in which selection likely acted on an introgressed locus, suggesting persistent interploidy gene flow has a local influence on patterns of genetic variation in *A. arenosa*.

INTRODUCTION

Whole-genome duplication (WGD) has occurred in many organisms across eukaryotic kingdoms and has profoundly shaped genome evolution (Kellis et al. 2004; Dehal and Boore 2005; Jiao et al. 2011). These large-scale genomic events are implicated in increased genomic complexity and are associated with adaptive radiations of major lineages throughout the tree of life (Dehal and Boore 2005; Jiao

et al. 2011). WGD is particularly frequent in plants; ancient WGD events are estimated to have occurred in 30-100% of angiosperm lineages (Stebbins 1950; Grant 1981; Masterson 1994; Cui et al. 2006). WGD is also implicated in speciation as it is one of the few processes that may instantaneously give rise to reproductive isolation due to the lower success of interploidy crosses, and may thus serve as a mechanism of sympatric speciation (Wood et al. 2009).

Many extant plant species are known to have multiple ploidy levels (see for review Ramsey and Schemske 1998; Soltis et al. 2010), showing that polyploidy remains an active force in plant evolution. It has also been suggested that the establishment of new autotetraploid populations may be affected by interploidy gene flow (Ramsey and Schemske 1998). Thus for species with ploidy variation, it is important to understand the dynamics of polyploid formation and establishment, interploidy gene flow, and the evolutionary timescale over which multiple ploidies coexist and shape species-wide patterns of genetic variation. New genomic approaches and methods to reconstruct polyploid history (Arnold et al. 2012) hold promise to enable new detailed understanding of evolutionary dynamics in multiploidy systems.

There are two major classes of polyploids: autopolyploids, which form from within-species WGD and generally randomly segregate homologs, and allopolyploids, which have a hybrid origin and usually diploid-like inheritance (Ramsey and Schemske 1998; Parisod et al. 2010; Bomblies and Madlung 2014). Of these, autopolyploids have received less attention in the evolutionary genetics literature, though in several species previous studies documented that autopolyploids arose multiple times and/or from more than one individual (Soltis et al. 1989; Brochmann and Elven 1992; Van Dijk and Bakx-Schotman 1997; Seagraves et al. 1999; Yamane et al. 2003; Yang et al. 2006; Luo et al. 2014). There is now good evidence that autopolyploids are more common than was previously appreciated (Soltis et al. 2010). Their often tetrasomic mating system (arising from random segregation of all four homologs) presents an intriguing problem and has important implications for autopolyploid population genetics.

Here, we reconstruct the evolutionary history of widespread autotetraploid populations of *Arabidopsis arenosa*. This species is newly being developed as a model for understanding the molecular basis of autopolyploid evolution (Hollister et al. 2012; Yant et al. 2013); for this it is particularly important to understand the evolutionary history of the polyploid lineage(s) and the degree of interploidy admixture in more detail. This species is an obligate outcrosser closely related to *A. lyrata* and the widely used model *A. thaliana* (Al-Shehbaz and O'Kane 2002), both of which have sequenced and annotated genomes (The Arabidopsis Initiative 2000; Hu et al. 2011). Autotetraploid *A. arenosa* has high genetic diversity (Hollister et al. 2012; Schmickl et al. 2012; Hohmann et al. 2014), and populations are widely distributed through much of Central and Northern Europe, while diploids are found in Eastern Europe and the Balkans, and along the southern Baltic Coast in Poland; the two types overlap extensively in the Carpathian Mountains (Schmickl et al. 2012; Kolar et al.2015). Both the Northern Carpathians and the Eastern unglaciated Alps are considered centers of diversity for the species and have evidence of genetic differentiation from one another, suggesting a degree of geographic isolation and perhaps independent origin of the resident tetraploids (Schmickl et al. 2012). Previous work also suggested that the tetraploids experience gene flow from diploids, but not the reverse (Jørgensen et al. 2011). However, the age of the polyploidy event(s) is not known, nor is it known whether polyploids arose once or have multiple origins.

We sampled populations of autotetraploid *A. arenosa* from across its Central European range, including the previously reported differentiated Northern Carpathians and Southeastern Alps regions (Schmickl et al. 2012), and several diploid regions from the Carpathian Mountains and adjacent Pannonian Basin. We use both parametric and nonparametric analyses of three distinct genomic sequencing datasets to infer the number and timing of autotetraploid origins of these samples, estimate the geographic location of origin(s), and quantify the extent and direction(s) of interploidy gene flow. We find that the geographically widespread populations we sampled of the autotetraploid *A. arenosa* likely arose from a single ancestral population, probably in the Northern Carpathians

approximately ~11,000-30,000 generations ago. Thereafter, the tetraploids split into at least four major lineages that colonized much of Europe, including both the Alpine and Carpathian regions, as well as flatlands of central Europe. We find evidence that geographically proximal diploid and tetraploid populations experienced past bidirectional interploidy admixture, and in rare cases, introgressed haplotypes may have come under selection in the recipient population. The methods we use and develop will be applicable in a wide range of species with ploidy variation.

RESULTS

Genomic Data and assessment of ploidy and inheritance

We collected seed samples from 6 diploid and 14 tetraploid *A. arenosa* populations from across a large swathe of its European range (Figure 1) and used three different genomic datasets (summarized in Table S1) to analyze their population history. First, we generated a Restriction-Associated DNA sequencing (RADseq; Peterson et al. 2012) dataset for 358 plants. We complemented this with two additional genomic datasets with overlapping population samples: (1) whole-genome sequencing of population pools (PoolSeq) which sampled 89 of these plants (Wright et al. 2015), and (2) a previously generated whole-genome sequencing dataset (IndSeq) that sampled a subset of 16 of these plants (Yant et al. 2013). The IndSeq dataset, though it samples fewer plants, serves as a standard, since it does not suffer from ascertainment biases potentially present in RADseq and PoolSeq data (Cutler and Jensen 2010; Arnold et al. 2013; Gautier et al. 2013) and has higher sequencing depths per chromosome, allowing more accurate single-nucleotide polymorphism (SNP) calls. We determined genotypes using the GATK (McKenna et al. 2010), which accommodates diploid and tetraploid samples, and only considered biallelic sites with sequencing depth cutoffs of 8 or higher per individual. All three datasets produce similar estimates of allele frequencies, though estimates of genetic diversity differ; relative to the IndSeq dataset, RADseq underestimates diversity, while PoolSeq overestimates it (Tables S2 and S3, Supplementary Methods). Both of

these follow expected trends based on previously identified biases in such datasets (Cutler and Jensen 2010; Arnold et al. 2013).

We previously assessed ploidy for several of these *A. arenosa* populations by flow cytometry (Hollister et al. 2012), but not all individuals were sampled. Therefore, we assessed the ploidy of each individual sampled here bioinformatically using the RADseq dataset and the simple logic that, at polymorphic sites, raw SNP count data for diploid and autotetraploid samples should be different. Specifically, the distribution of non-reference base counts for all polymorphic sites within a single diploid individual should resemble a binomial distribution with a mean of 0.5, while the same distribution for an autotetraploid should be trimodal, due to an amalgamation of three distinct binomial distributions with means of 0.25, 0.5, and 0.75, as autotetraploids have three types of heterozygotes. Using the RADseq dataset, limiting ourselves to filtered heterozygous sites within an individual that have sequencing depths of at least 30, counting the number of non-reference base calls and comparing them to simulated expected distributions using a *G* statistic (see Materials and Methods), we easily discriminated between samples of different ploidy (Figure S1, Table S4). With the exception of one putatively tetraploid individual found in an otherwise diploid population (and excluded from subsequent demographic analyses), all samples from a collection site were of the same ploidy. This is consistent with previous findings in *A. arenosa* (Schmickl et al. 2012; Kolar et al. 2015). The one tetraploid we found in a diploid population is a potentially spontaneous neotetraploid in population D5 (as opposed to a migrant from a tetraploid population), as it is genetically similar to diploids sampled from the same population (Figure S2). Although two additional diploid samples did not have simple binomial non-reference base count distributions, they were likely diploid (see Materials and Methods).

An important assumption when modeling autotetraploid data using the coalescent is that chromosomes are exchangeable (Arnold et al. 2012). This assumption would be violated if there were restricted recombination to certain chromosome pairs, as would occur if chromosomes have pairing partner preferences. Structure between duplicated chromosome sets due to pairing

preferences should create an enrichment of alleles at 50% frequency relative to an allele frequency spectrum (AFS) of a population with unstructured chromosomes (Hollister et al. 2012). Neither the IndSeq nor the RADSeq datasets display an excess of alleles at 50% frequency (Figure S3A), and tetraploid genotype proportions closely resemble those expected under Hardy-Weinberg equilibrium for tetrasomic inheritance (Figure S3B). These data confirm that the assumption of random chromosome assortment is not violated and that *A. arenosa* populations retain fully tetrasomic inheritance (as previously shown for a smaller set of samples in Hollister et al. (2012)).

Principal component analysis suggests a single geographic origin of the sampled tetraploids

To study the genetic relatedness of sampled *A. arenosa* populations, we used principal component analysis (PCA), a non-parametric approach that allows for multiple ploidies. For this analysis we used the RADseq dataset, which included the largest number of individuals sampled. Since PCA is sensitive to sample sizes (Novembre and Stephens 2008), we used a subsampling approach to control for the disparity in diploid and tetraploid representation within the RADseq dataset. A PCA of only diploids (10 individuals per population, 11,758 SNPs) shows there are two distinct groups within our diploid samples (Figure 2A), one found in the Carpathian Mountains in Romania and Slovakia (D1-D3) and another in the biogeographically distinct Pannonian Basin in Southern Slovakia and Central Hungary (D4-D6). To study the relationship of the tetraploids we sampled with these diploid gene pools, we added a subsample of 30 tetraploids from across the *A. arenosa* range (6,117 SNPs, Figure 2B) and found that all tetraploids group with Northern Carpathian diploids (D2 and D3).

These PCA results contain two important pieces of information: (1) the tetraploids, despite their high genetic diversity and geographically widespread distribution, are all comparatively closely related to one another, and (2) all the tetraploids we sampled are most closely related to the Northern Carpathian diploids in our sample. To better visualize the relationship between the tetraploids and their

closest sampled diploid relatives, we performed another PCA using just Carpathian diploids (D1-D3). To circumvent the large differences in sample sizes between diploids and tetraploids, we used 30 diploids (10 per population) and a single tetraploid individual to elucidate how each tetraploid relates to the principal component space of diploid genetic variation. We repeated this analysis 140 times, sampling 10 individuals from each of 14 tetraploid populations, and superimposed the results onto the same pair of principal component axes (415,718 SNPs, Figure 2C). The tetraploids cluster together between the three diploid populations, suggesting there is a single tetraploid gene pool within our sample (which was sampled broadly across the tetraploid *A. arenosa* range). As before, the tetraploid gene pool is more closely related to the Northern Carpathian diploids than the Southern Carpathian diploids we sampled. Three tetraploid populations radiate from the central cluster towards each of the three diploid populations, suggesting there may have been admixture. In each case, the populations where admixture is suggested are the most geographically proximal to each respective diploid. We further explore the possibility of admixture below using coalescent analyses. A single PCA with these 30 Carpathian diploids and 30 tetraploids from across the range yields similar results (Figure S4).

Demographic modeling affirms a single geographic tetraploid origin with ancient interploidy admixture

The above analyses suggested the 14 widespread autotetraploid populations arose from a single ancestral population closely related to extant Northern Carpathian diploids, with potential admixture among geographically proximal diploid and tetraploid populations. To verify this result and quantify admixture proportions, we explicitly modeled the history of these populations using the coalescent. We modeled groups of three populations in each case, in an approach we call "trio analyses." In each analysis, we constructed models of one diploid and two tetraploid populations, and tested which of two models better fits observed polymorphism data: (1) a single tetraploid origin allowing for subsequent interploidy admixture between geographically proximal populations (model A in

Figure 3) or (2) two independent tetraploid origins with potential admixture between tetraploid populations (model B in Figure 3). While it would be possible to generate increasingly complex models, limiting analyses to trios avoids the problem of excess empty categories in the multidimensional allele frequency spectrum (AFS) that would occur if more populations were included in each analysis.

For each trio analysis, we always included one tetraploid that is geographically distant from diploids and displayed no evidence of interploidy admixture according to simple demographic models (i.e. T7 and T13, Figure S5). In addition to a one-time, bidirectional admixture event in which populations are allowed to potentially exchange a larger proportion of genetic lineages, low levels of equilibrium migration among demes is also allowed. For these analyses, we used 4-fold degenerate sites (coding sites where mutations to any base will not alter amino acid sequence) of populations from our RADseq dataset, since these data produced similar model parameter estimates as the IndSeq data (Table S5, See Materials and Methods below).

We find that among our samples the single tetraploid origin model is unambiguously supported in all trio analyses (Table S6). In cases involving geographically proximal tetraploids and diploids, we also find evidence of subsequent interploidy admixture. For example, tetraploids T1, T2, T5, and the railway tetraploids (T4, T6, T14) share more genetic variation with Carpathian diploid populations than tetraploid populations further within the tetraploid range (i.e. T7 and T13; Figure 2C, Figure S5). Nevertheless, population trio analyses show these admixed tetraploid populations diverged from the same ancestral population as T7 and T13 (Table S6A,B), and that allele sharing thus reflects subsequent admixture, not independent origins. Analyses of the same populations present in the PoolSeq dataset validated these results (Table S7).

Unidirectional gene flow from diploid to tetraploid *A. arenosa* was previously reported (Jørgensen et al. 2011). However, in our models, maximum likelihood estimates (MLEs) of bidirectional admixture proportions strongly suggest that interploidy gene flow occurred among geographically proximal populations in both directions (Table 1). This result holds for 4-fold degenerate as well as noncoding

sites (Table S8) and for both RADseq and PoolSeq datasets. Moreover, this result is robust to higher sequencing depths and thus greater genotype-calling accuracy (Table S9). Similar models that only allow for unidirectional admixture from diploids to tetraploids invariably have significantly lower likelihoods than those that allow bidirectional admixture (Table S10).

Interploidy admixture appears to be a local effect, as we estimated admixture proportions using tetraploid populations that are geographically distant from any diploid and found these are near zero (Table S11). For these populations, models that do not allow for any interploidy admixture fit the data significantly better than those that do allow admixture (Table S10B). The PoolSeq data again gives the same results (Table S10C). Major interploidy admixture events (as opposed to ongoing background levels of admixture) may also be relatively ancient; for all trio analyses involving populations with interploidy admixture, the 95% parametric bootstrap CIs for the timing (in generations) of large-scale admixture events overlap with the CIs for the divergence time of the two tetraploid populations in the trio (Table 1).

The ancestral tetraploid is most closely related to Northern Carpathian diploids

We modified the trio analysis used above to confirm the likely geographic origin of the autotetraploid lineage we sampled. Here, we used two diploid and one tetraploid population in each analysis to test whether tetraploids are more closely related to a particular diploid gene pool within our sample, while again accounting for interploidy admixture when present (Figure S6). Since we already established above that the tetraploids are more closely related to Carpathian diploids (D1-D3) than to Pannonian diploids (D4-D6, Figure 2), we only used Carpathian diploid populations. When the Romanian diploid (D1) is included, the tetraploid population used is consistently more closely related to one of the Slovakian diploid populations (either D2 or D3; Table S12). This explicit modeling of population history agrees with results from PCA (Figure 2B) and suggests that the ancestral population from which the widespread tetraploid lineage originated is derived from a diploid lineage whose closest extant relatives are found in the Northern Carpathian Mountains

today. This area corresponds to what Schmickl et al. (2012) called the "cradle of speciation" for the *A. arenosa* species complex.

Age of the tetraploid lineage

To estimate the age of this tetraploid lineage, we constructed a model for coalescent analyses using a population from the oldest tetraploid split (T1 in Romania) and the closest diploid relative we sampled (D3 in Slovakia). The estimate of the oldest tetraploid divergence in our sample serves as a lower bound to the age of the tetraploid, assuming this split occurred soon after the ancestral tetraploid arose in the Northern Carpathians. Likewise, the estimate of the divergence between the ancestral tetraploid and the closest diploid relative serves as an upper bound to the age of the tetraploid. We estimated these divergence times by constructing a model of four populations (T1,T5,D1,D3) that accounts for interploidy admixture (Figure 4). The MLE for the divergence between the Romanian and Slovakian tetraploids (T1 and T5) is ~15,000 generations, while the MLE for the divergence between the ancestral tetraploid and Slovakian diploid D3 is ~19,000 generations. Thus, the ancestral population that gave rise to the widespread tetraploid arose ~15,000-19,000 generations ago (or ~11,000-28,000 generations using the 95% parametric bootstrap confidence intervals, Figure 4).

Genetic structure within tetraploids reveals distinct clades and multiple migration routes

To better understand the demographic history of the tetraploid lineage after it arose, we conducted three separate analyses using only tetraploids: STRUCTURE, *Treemix*, and PCA (Figure 5). STRUCTURE results indicate that tetraploid populations in our sample fall into five major clades that roughly correspond to geographic origin. The exception is that samples collected from railroads defy the general trend of isolation-by-distance (Figure 5A). Population graph analysis with *Treemix*, a program that constructs a population tree and allows admixture, supports the STRUCTURE results of five tetraploid clades, although bootstrap support for a few nodes are low (Figure 5B). We also performed coalescent analyses

using one population from each of the four non-railroad clades, and this showed that the tree topology in Figure 5B fits the data significantly better than all other possible topologies (Table S14). PCA broadly agrees with STRUCTURE and *Treemix* results (Figure 5C), and the patterning of individuals within the first two principal components resembles the null expectation of a stepping-stone model across Europe (Novembre and Stephens 2008). However, principal component three (Figure 5D) highlights differences between Southwest German (Swabian) and Alpine clades. All three analyses in Figure 5 show that the geographically diffuse, panmictic network of railroad tetraploids has admixed with a population from the Alpine clade, T11, which grows on a railway in the Alps. Nearby population T10 was also collected near a railway and exhibits low levels of admixture (Figure 5A). Although *Treemix* results suggest admixture between T1 and the railroad tetraploids, this may be an artifact of not including D1 with which they have admixed (Table 1, Figure 2C, Figure S4).

From *Treemix* analysis (Figure 5B) and corroborating coalescent simulations, we can infer the oldest split within our sample separates the Southern Carpathian tetraploid from other populations, suggesting an early migration event likely along the Carpathian Mountains into Romania (T1), while the second oldest split involved a lineage that we sampled from the Swabian Alb in Southwestern Germany (T12, T13). Tetraploids sampled from the Alps (mostly in Austria) are more closely related to Slovakian tetraploids from the Northern Carpathian Mountains (Figure 5B), suggesting they may represent a single colonization route along the Carpathian mountains into the Alps. The PCA in Figure 5D agrees with the interpretation that the Swabian and Alpine lineages represent separate radiations out of an ancestral Slovakian clade.

Interploidy admixture introduced alleles that came under selection

Among several populations, we found evidence of bidirectional admixture. This may increase levels of genetic variation in both ploidies and raises the possibility that gene flow could introduce adaptive alleles. We thus looked for regions of the genome in which proximal diploid and tetraploid populations have

experienced selection on the same set of genetic variants where these were likely transferred by gene flow (rather than representing shared ancestral variation). We identified candidate events using admixed populations in the PoolSeq dataset. We scanned both admixed population pairs for regions in which both ploidies displayed evidence of selection on similar sets of geographically unique SNPs. Specifically, we identified loci with significantly low values of Fay and Wu's H (Fay and Wu 2000) that also display an excess of high-frequency, geographically unique shared variants compared to genome-wide patterns. We required that both populations display evidence of selection because we do not know the direction of admixture for particular loci. Furthermore, finding haplotypes under positive selection in one population, but not in the other, could also be explained by selection preceding neutral gene flow. This method is thus conservative and will not detect all loci that may have experienced selection after admixture.

We find only two examples where there is evidence by our criteria of selective events following admixture in likely introgressed genomic regions. These signals are not due to fluctuations in sequencing depth, as local depths are similar to genome-wide averages (Figure S7). First, in populations D1 and T1, only one region has 5% outlier low values of Fay and Wu's H in both populations, indicating an excess of high-frequency derived variants (Figure 6A). Using only shared SNPs unique to D1 and T1 relative to all other sampled populations, we calculated θ_H, a metric sensitive to high-frequency derived variants (Fu 1995). Elevated θ_H in this genomic region suggests it is enriched for geographically unique, high frequency shared variants in both populations (Figure 6B). These polymorphisms are closely linked (Figure 6C,D). This metric contrasts with the version of θ_H used in Fay and Wu's H , which uses all SNPs. We find no evidence of selection in other populations at this locus, suggesting the allele found in these two populations may be locally adaptive (Figure S8). Two genes within this region have many high-frequency derived SNPs (relative to the *A. lyrata* reference and other *A. arenosa* populations) that are shared among D1 and T1, one of which causes an amino acid change (Table S15). The two genes are orthologs of *A. thaliana* genes AT3G63330 and AT3G63340,

both of which encode protein phosphatases of otherwise unknown function (Table S16).

Second, we found a single genomic region distinct from the one identified in D1 and T1 in which admixture may have been followed by positive selection in populations D3 and T5 (Figure S9). Although selection is not apparent in most other populations at this locus, it may be occurring in a second geographically proximal diploid population D2 (Figure S10) that has admixed with T2 (Table 1). This region contains a single gene with numerous high-frequency shared SNPs, one nonsynonymous, in D3 and T5. These polymorphisms are absent from other tetraploid populations, but are also present in D2 (Table S17). The single gene with high frequency derived polymorphisms in D3 and T5 encodes a protein with pollen allergen domains that is highly homologous to three *A. thaliana* genes in the β-expansin family (AT2G45110, AT1G65680, AT1G65681, Table S18). The proteins encoded by these genes are involved in loosening of plant cell walls e.g. during the penetration of pollen tubes through the stigma and style during sexual reproduction (Cosgrove et al. 1997). Expansins may also be important for cell growth in polyploids, and two expansins, including AT1G65680, were found to be under selection in polyploid *A. arenosa* in our previous work (Yant et al. 2013).

DISCUSSION

Here, we analyze three genomic datasets to assess the demographic history of 14 autotetraploid *A. arenosa* populations we sampled broadly from the Central European portion of its range, and its diploid relatives. We find that the 14 autotetraploid populations comprise a single lineage that likely radiated from a single ancestral population ~11,000 - 30,000 generations ago, and is closely related to populations found in the Northern Carpathians today. Since *A. arenosa* is generally perennial (Al-Shehbaz and O'Kane 2002), but flowers every year, with railway populations biennial or annual (K Bomblies, P Baduel & B Hunter, unpublished), each generation likely corresponds to one or two years. These results extend previous work on *A. arenosa*, which suggested the Carpathian Mountains as a

center of diversity for the species (Schmickl et al. 2012; Hohmann et al. 2014). Work on other autotetraploids has shown that autotetraploid lineages often arise from more than a single individual (Soltis et al. 1989; Brochmann and Elven 1992; Van Dijk and Bakx-Schotman 1997; Seagraves et al. 1999; Yamane et al. 2003; Yang et al. 2006; Luo et al. 2014), showing that WGD may be an ongoing mutational process. For *A. arenosa*, the ancestral autotetraploid population also likely arose from multiple neopolyploid individuals (since *A. arenosa* tetraploids are highly diverse and obligately outcrossing). Unreduced gamete formation in diploids, perhaps elevated in cold stress conditions during periods of glaciation, likely played an important role in the formation of this ancestral gene pool (Ramsey and Schemske 1998). We note, however, that we cannot conclude that all extant *A. arenosa* tetraploids originated from this single ancestral population. There are additional tetraploid populations in regions we have not sampled that may have independent origins (Schmickl et al. 2012; Kolar et al. 2015). Our finding of an apparently new tetraploid in a Pannonian diploid population supports the idea that tetraploids can occasionally arise spontaneously in *A. arenosa*. Our analyses do allow us to rule out that a more distantly related diploid than the Northern Carpathian diploids gave rise independently to any of the tetraploids within our sample.

Polyploidy is unusual in that it can immediately present a strong gene flow barrier between ploidies even in sympatry, due to the low fertility of progeny from interploidy crosses (e.g. triploids; Ramsey and Schemske 1998). Nevertheless, the autopolyploidization of *A. arenosa* did not create immediate complete reproductive isolation, as our parametric and nonparametric analyses detect multiple, independent cases of interploidy admixture between geographically proximal populations. Our reconstructions of evolutionary history show this admixture was likely extensive, ancient, and importantly, bidirectional. This is in contrast to a previous report for *A. arenosa*, which suggested that gene flow had occurred only from diploids to tetraploids, not the reverse (Jørgensen et al. 2011). Bidirectional gene flow among ploidies is, however, consistent with findings from other plant species (Thórsson et al. 2001; Ståhlberg 2009).

Gene flow from diploids to tetraploids can occur without the formation of triploids, since diploids produce unreduced gametes at some frequency that can fertilize tetraploids, or neo-tetraploids can arise spontaneously that can also fertilize established tetraploids (Ramsey and Schemske 1998). Our observation of an apparently newly formed tetraploid in a diploid population supports the possibility that the latter mechanism can occur in *A. arenosa*. Gene flow from tetraploids to diploids, on the other hand, necessitates the formation of triploids, since there is no known mechanism of nondisjunction by which tetraploids can make haploid gametes to regenerate diploids. Though triploids have low fertility, they generally do retain some fertility, allowing gene flow to occur via a so-called "triploid bridge" (Ramsey and Schemske 1998). That this is possible in the *Arabidopsis* genus is supported by the observation that triploids in *A. thaliana* can generate viable aneuploids and populations ultimately resolve to stable diploids and tetraploids over several generations of selfing (Henry et al. 2005). We did not identify any triploids in our sampling of 358 plants, but previous studies have observed rare triploids in *A. arenosa* (Kolník 2007; Jørgensen et al. 2011; Kolar et al. 2015), which may suffice to yield substantial gene flow over evolutionary timescales. Since estimates of interploidy gene flow tend to be older than several thousand generations in our models (Table 1), it is possible that some degree of interploidy reproductive isolation has evolved and that triploids were once more abundant than they are now.

What the consequences are for diploids of the influx of tetraploid alleles or vice versa is not known. We speculate that interploidy admixture, while generally neutral or likely at times deleterious, could occasionally result in the exchange of beneficial haplotypes. That introgressed regions are beneficial and subsequently experience positive selection seems to be rare, but we do find two cases where admixed populations seem to have experienced selection in introgressed genomic regions (Figure 6, S9). In both cases, there is evidence of a strong selective sweep in the admixed diploid and a weaker signature of selection in the corresponding tetraploid. This pattern may be explained by autotetraploids generally having weaker responses to selection (Hill 1971), from selection taking place further in the

past in the tetraploid, or both. While these results may also be explained by parallel selection on haplotypes segregating in both ploidies as standing genetic variation, we do not think this is likely as the hitchhiking effect is stronger than expected if the selected SNP(s) persisted as neutral variant(s) as long as the divergence time between D1 and T1 (~35,000 generations). Parallel selection on standing variation this old would likely produce a softer sweep undetectable by Fay and Wu's H (Messer and Petrov 2013). Ultimately, having haplotype information for this region would help resolve this uncertainty. It is also possible these loci are not adaptive, but experienced selection upon introgression due to other factors such as meiotic drive (Derome et al. 2004).

After the ancestral tetraploid population arose from its diploid progenitor, it colonized much of Europe via at least four distinct migration routes from its likely origin in the Northern Carpathian Mountains. One lineage is represented in our sample by a single population from the Southern Carpathians that diverged from other tetraploids ~12,000 generations ago (S. Carpathian clade, Figure 5). This is the oldest tetraploid divergence time in our sample, and this colonization may have been possible from large ice-free swaths within the Carpathians, even during the last glacial maximum (reviewed in Ronikier 2011). At least two tetraploid lineages then independently colonized southwest Germany and the Alps, diverging from each other ~8,000 generations ago. The dating of these events strongly depends on the mutation rate we estimated from the data (see Materials and Methods) and used in coalescent analyses, but these dates can simply be rescaled if a different mutation rate is discovered for *A. arenosa*.

The migration route of the populations found currently in the Southwestern German Swabian Alb region is unclear and may have occurred along the chains of limestone hills that run across Germany north of the Alps. Finally, a fourth lineage liberated itself from the generally montane niche that other tetraploid lineages are found in and colonized railroad habitats across Central and Northern Europe. This genetically and phenotypically distinct "railroad ecotype" has rapidly traversed large geographic distances such that populations sampled from disparate parts of the range remain very similar, which is not true of the other tetraploid lineages. This

suggests a rapid and recent range expansion, likely facilitated by migration along railway networks. To this last point are clearly a few exceptions: we sampled one population from a railway in the Alps (T11) that is genetically primarily an Alpine type. This second colonization of railway habitats may have been facilitated by admixture with the more prevalent railway ecotype found in other parts of Europe (Figure 5).

In sum, we show that *A. arenosa* autotetraploids sampled from 14 widely distributed populations all originated from a single ancestral population that likely arose ~11,000 - 30,000 generations ago in the Northern Carpathian Mountains. After whole genome duplication, this polyploid population subsequently split into at least four distinct lineages that colonized the Southern Carpathians, Southwestern Germany, the Alps, and the railways of Central and Northern Europe. We also show evidence that there has been bidirectional interploidy admixture among geographically proximal diploid and tetraploid populations. Tetraploids that colonized the Alpine region, where no diploids occur, show no evidence of past interploidy admixture. In two instances gene flow between diploids and tetraploids exchanged sets of variants that are associated with selection in both ploidies, suggesting that bidirectional admixture may have functional consequences, though whether the alleles that came under selection are adaptive remains to be tested. Nevertheless, these results suggest that interploidy admixture within multiple-ploidy systems may shape patterns of variation. Our recovery of an apparently newly formed tetraploid individual in a population of diploids suggests that polyploids do arise sporadically within *A. arenosa* diploid populations, and may give rise to independent tetraploid lineages; some of these may persist, though they were not represented in our sampling. The occasional formation of neotetraploids could provide an additional mechanism for gene flow from diploids to tetraploids, and has at least the theoretical potential to generate novel tetraploid lineages.

MATERIALS AND METHODS

Generation of DNA sequence data

We used three DNA sequence datasets in this analysis: Restriction-associated DNA sequencing (RADseq), individual whole-genome sequences (IndSeq), and whole-genome sequencing of population pools (PoolSeq, Table S1). The generation of the IndSeq and PoolSeq datasets was described previously (Yant et al. 2013; Wright et al. 2015). We generated the RADseq dataset using a modified a double digest RADseq protocol (Supplementary Methods; Peterson et al. 2012).

DNA sequence alignment and variant calling

For all datasets, we aligned DNA sequences to the *Arabidopsis lyrata* reference genome (Hu et al. 2011) using Stampy v1.0.21 (Lunter and Goodson 2011) with default parameter values. For the IndSeq and PoolSeq datasets, we removed PCR duplicates using Picard (http://picard.sourceforge.net/). We locally realigned reads around indels for all datasets using the Genome Analysis Toolkit (GATK v2.7; McKenna et al. 2010). We called sequence variants for the IndSeq and RADseq datasets using the GATK. To maximize variant detection within and between populations for each dataset, we genotyped all individuals irrespective of ploidy simultaneously as diploid or tetraploid, with individual genotypes later extracted from the appropriate file. Potential variants were filtered using the GATK VariantFiltration tool (Supplementary Methods). To call variants in the PoolSeq dataset, we used SNAPE (Raineri et al. 2012, as described in Wright et al. 2015). We also used an additional data filtration step to remove loci that likely contain spuriously mapped sequence reads (Supplementary Methods).

Bioinformatic assessment of ploidy

To determine the ploidy of each sample, we compared non-reference base count distributions of each sample to those expected of a diploid, triploid, and tetraploid. We simulated the expected distribution for diploids as Binomial(n=30, p=1/2), since we only considered sites with a minimum sequencing depth of 30. To model triploids, we constructed a compound distribution in which non-reference base counts were simulated from either Binomial(n=30, p=1/3) or Binomial(n=30, p=2/3) with probability 2/3 or 1/3, respectively, as expected for a neutral mutation

frequency spectrum (Fu 1995). For tetraploids, we simulated non-reference base counts from Binomial(n=30, p=1/4), Binomial(n=30, p=1/2), or Binomial(n=30, p=3/4) with probability 6/11, 3/11, or 2/11, respectively. We compared observed non-reference base counts of each sample to these simulated expected distributions using a G statistic in which $G = 2\sum_i O_i * \ln(\frac{O_i}{E_i})$, where O_i are observed frequencies and E_i are expected frequencies. For calculating G, we categorized alternate base count proportions greater than 0.2 and less than 0.8 into twelve bins (increments of 0.05) to avoid sequencing errors that occur at low frequencies. Using G to select the best-fit model for each sample, we show all samples from a collection site were of the same ploidy (Table S4) with three exceptions: a putative neotetraploid in an otherwise diploid population (Figure S2), and two samples that were likely diploid but not well-modeled by our expected diploid distribution due to greater, unexplained variance in base count frequencies (Figure S11).

Principal Component Analysis

We performed principal component analysis (PCA) in R using the package *adegenet* 1.3-5 (Jombart and Ahmed 2011), which accommodates for variable ploidy. All PCAs used SNPs in which all individuals had a sequencing depth of at least 8x. For the single tetraploid PCA, we allowed up to 30% of individuals to have a sequencing depth of less than 8x and coded that site as missing for those individuals.

Coalescent analyses

We fit demographic models in population trio analyses to observed data using *fastsimcoal2* (Excoffier et al. 2013), a program that uses the coalescent (Kingman 1982) to simulate multi-dimensional AFS and a modified expectation-maximization algorithm to search parameter space and find maximum likelihood estimates (MLEs) for model parameters. After MLEs are obtained, we compared model likelihoods with Akaike information criterion (AIC) to assess which model had a higher probability of being correct given the candidate set of models. To avoid an excess of zeroes in higher-dimensional AFS used for analysis, we used only three

populations and sampled only 6 tetraploids and 9-12 diploids for each population. We only considered sites in which all individuals had a sequencing depth of at least 8x, using the common allele of 24 *A. lyrata* genomes as the reference allele.

Since demographic analyses are potentially sensitive to an enrichment of high-frequency alleles due to misspecification of the ancestral allele, we attempted to correct the three-dimensional AFS for mispolarized alleles using the following extension of the technique described in Baudry and Depaulis (2003). Although we excluded sites with more than two segregating bases from analyses, multiple mutations may occur at a site and go undetected if the same mutation occurs twice within the species tree. We calculated the probability of a biallelic site experiencing multiple mutations using the proportion of triallelic sites in the sample, empirically derived estimates of transition and transversion rates, and equation 3 in Baudry and Depaulis (2003). After an uncorrected, unfolded 3D AFS was obtained from the data, we constructed a folded 3D AFS. We multiplied each entry in the folded 3D AFS by the empirically obtained probability of a biallelic site experiencing more than one mutation to obtain frequency-specific proportions of mispolarized alleles. We used these proportions to reorient a proportional number of alleles in the respective unfolded category (Figure S12).

To construct 95% parametric bootstrap confidence intervals (CIs), we simulated the same number of sites used in the analysis 100 times, using linkage blocks of 260bp (insert size) for the RADseq dataset or 5kb for the IndSeq and PoolSeq datasets, and a population recombination rate that is roughly twice as large as the population mutation rate. We estimated the mean population recombination rate per bp using *LDhat* (Auton and McVean 2007) on diploid whole-genome sequences, specifically from an 800kb segment on chromosome 2 in four individuals from population D3. We chose this chromosomal segment due to even and high sequencing depths. For each simulated dataset, we ran 50 instances of *fastsimcoal2* to infer the MLEs of parameter values, which were then used to construct confidence intervals.

In order to obtain absolute values for model parameters, a mutation rate must be specified. We calculated the mutation rate for noncoding and 4fold-

degenerate sites using a simple isolation-migration (IM) model with populations D2 and D3. We obtained 100 MLEs of all parameter values, including the mutation rate, using 50 *fastsimcoal2* runs each time to obtain parameter estimates. We repeated this analysis separately for noncoding and 4fold-degenerate sites. The mode of the 100 MLEs for the mutation rate was assumed to be near the true mutation rate, since this held true for 100 datasets simulated with a known mutation rate (Figure S13). We thus used a mutation rate of 3.7×10^{-8} and 4.3×10^{-8} for noncoding and 4fold-degenerate sites, respectively.

Comparison of demographic inference among genomic datasets.

In order to evaluate the sensitivity of demographic inference to dataset type, we generated a simple IM model for diploid populations D2 and D3 for all three datasets and calculated the maximum likelihood estimates (MLEs) of model parameters using *fastsimcoal2* (Excoffier et al. 2013). IndSeq and RADseq datasets produced very similar MLEs of parameter values for divergence time and comparable migration rates when only 4-fold degenerate sites (sites which can sustain any mutation without causing an amino acid change) were used (Table S5). Thus we used this functional category of sites for coalescent-based reconstructions of history; the use of noncoding sites caused results to differ more significantly between datasets (Table S5).

Reconstruction of tetraploid history

We characterized tetraploid population structure using STRUCTURE v2.3.4 (Pritchard et al. 2000), selecting a value of K populations that corresponded to the last largest increase in likelihood before likelihood values approached an asymptote with increasing values of K. We constructed population graphs of the tetraploids with *Treemix* (Pickrell and Pritchard 2012), using population T1 as root and adding migration edges until residuals did not appreciably decrease (two in our case). We bootstrapped the data by generating 1000 replicates, subsampling every three SNPs each time, and using Newick Utilities (Junier and Zdobnov 2010) to summarize results. Noncoding sites were used for *Treemix* analysis.

Acknowledgements

We thank John Wakeley, David Reich, Kevin Wright and Russ Corbett-Detig for helpful advice and discussions. We also thank Pierre Baduel for helping design the map used in Figure 1. This work was funded by an NSF predoctoral fellowship to BA (2010095484) an NSF DIGG to BA (1210308) and an NSF grant to KB (IOS-1146465). The funders played no role in study design or interpretation.

Table 1. MLEs of model parameters using population trios.

	Population Trio			
Parameter	**D3,T5,T7** 63,669 sites	**D1,T1,T7** 93,914 sites	**D2,T2,T7** 75,373 sites	**D1,T4,T7** 82,122 sites
Adm_{DT}	**0.22** (0.08, 0.41)	**0.25** (0.12, 0.39)	**0.33** (0.13, 0.44)	**0.24** (0.09, 0.34)
Adm_{TD}	**0.44** (0.25, 0.54)	**0.09** (0.06, 0.19)	**0.22** (0.14, 0.35)	**0.10** (0.02, 0.18)
T_{Adm}	**6877** (4546, 9153)	**9317** (7354, 12162)	**5023** (4129, 8439)	**6995** (4253, 9920)
D_1	**8318** (6100, 12106)	**12629** (9312, 16276)	**7077** (5731, 11076)	**7814** (4762, 11674)
D_2	**33837** (20046, 37053)	**55298** (35567, 68114)	**40911** (30388, 70942)	**49139** (30707, 61092)
N_D	**79142** (52419, 93859)	**60119** (44793, 72179)	**43930** (34351, 65234)	**50600** (38264, 66971)
N_T	**45256** (33292, 59610)	**118047** (86096, 137344)	**48340** (37030, 76297)	**42253** (27038, 64637)
$N_{T'}$	**96011** (70612, 122310)	**96725** (71698, 121498)	**77614** (59388, 122849)	**59750** (38873, 92965)

Table 1 Notes: MLEs were obtained using the RADseq dataset, with 95% parametric bootstrap CIs shown below each number. Shown are the admixture proportions from diploids to tetraploids (Adm_{DT}) and tetraploids to diploids (Adm_{TD}) going backwards in time, the time of admixture (T_{Adm}), the divergence time between tetraploids (D_1) and the ancestral tetraploid and diploid (D_2), and the population sizes of the diploid (N_D), admixed tetraploid (N_T), and the outgroup tetraploid ($N_{T'}$). Divergence times are expressed in generations, population sizes are

in haploid number of chromosomes, and the number of sites used in each analysis is listed below the trio.

Figure 1. Map of central Europe with sample collection sites. Diploids are labeled D1 through D6 and colored in hues of red to yellow. Tetraploids are labeled T1 through T14 and colored in hues of light blue to purple. Both number and coloring schemes correspond to a longitudinal gradient from East to West. Tetraploids collected from railway habitats are labeled with X's. Our sampling includes collection sites within Romania (D1, T1), Hungary (D5, D6), Slovakia (T2, T3, T5, D2, D3), Poland (T4, T6), Austria (T7-T10), and Germany (T11-T14).

Figure 2. PCA of diploid and tetraploid *A. arenosa*. (**A**) PCA of all diploid populations separates groups D1-D3 and D4-D6 on PC1. PC2 primarily corresponds with latitude. (**B**) PCA as in (**A**), but including a subsample of tetraploids. Axes in (**A**) and (**B**) are labeled with the proportion of the total variance explained by that principal component. (**C**) PCA of single tetraploid individuals with diploid populations D1-D3. A separate PCA was performed for each tetraploid, 10 individuals from each of 14 tetraploid populations, and superimposed onto the same PC axes. Populations T1, T2, and T5, which have admixed with populations D1, D2, and D3, respectively, are labeled to show how they radiate out from the tetraploid group towards the diploid population with which they have exchanged alleles. Axes are labeled with the range of the percent of the total variance explained by that principal component across the 140 PCAs represented.

Figure 3. Models of population trios used to infer the number of times the tetraploid arose from a diploid ancestor. A present-day tetraploid population (T) may share more genetic variation with a geographically proximal diploid population (D) than a second, geographically distant tetraploid population (T') because of interploidy admixture (Model A) or because of a second, independent origin of a tetraploid population from the diploid ancestor (Model B). In addition to

equilibrium migration rates between all populations, a one-time, bidirectional admixture event was allowed (black arrows).

Figure 4. Coalescent model to estimate the age of the tetraploid. Divergence times (in generations) are shown for the oldest tetraploid split (T5 and T1) and for the split between the ancestral tetraploid and its potential diploid progenitor (D3 and the ancestor of T5 and T1). 95% parametric bootstrap confidence intervals of divergence times are listed below estimates in parentheses. This model accounts for interploidy admixture between geographically close diploids and tetraploids (bidirectional black arrows among N. Carpathian and S. Carpathian populations).

Figure 5. Genetic structure within the tetraploid. (**A**) STRUCTURE groups tetraploids into five major clades that correspond to geographic regions, with the exception of tetraploids collected from railroads, which cluster together irrespective of geography. There has been extensive admixture between these railway tetraploids and a population from the Alpine clade, T11. (**B**) Population graph analysis with *Treemix* supports STRUCTURE results of five clades with admixture and reveals the evolutionary relationships among clades. Migration edges (red arrows) indicate evidence for admixture, with numbers indicating migrant ancestry percentages. Bootstrap values under 90% are shown. (**C**) PCA using the first two principal components shows individuals cluster according to geographic region (colored according to STRUCTURE results) with the exception of the admixed T11. (**D**) The third principal component reveals the genetic structure between the Swabian (SW German) and Alpine clades. PCA axes are labeled with percent of the total variance explained by that principal component.

Figure 6. Evidence for selection following admixture. (**A**) Fay and Wu's H for 100 SNP windows is below the 5% quantile (dashed line) for both D1 (gray) and T1 (black) near the end of chromosome 5. (**B**) θ_H calculated in 50 SNP windows, using only shared variation unique to D1 and T1. Both D1 (gray) and T1 (black) have an

excess of high frequency-shared variation. The dashed lines show the genome-wide 95% quantile for each distribution. (**C**,**D**) Allele frequency plots for the region spanning the thick black line on the x-axis in (**A**) and (**B**) show a strong enrichment of geographically-unique, high-frequency shared variation for D1 (**C**) and to a lesser degree for T1 (**D**). In (**C**), alleles in D1 that are also present in T1 but absent from other Carpathian diploids (D2, D3) are shown. In (**D**), alleles in T1 are shown if present in D1, but absent from other tetraploids (T5, T7). Many of these variants fall within genic regions (black lines above **C** and **D**).

CITATIONS

Al-Shehbaz IA, O'Kane SL. 2002. Taxonomy and phylogeny of *Arabidopsis* (Brassicaceae). The Arabidopsis Book, 1: e0001.

Arnold B, Bomblies K, Wakeley J. 2012. Extending coalescent theory to autotetraploids. Genetics 192: 195-204.

Arnold B, Corbett-Detig R, Hartl D, Bomblies K. 2013. RADseq underestimates diversity and introduces genealogical biases due to nonrandom haplotype sampling. Mol Ecol. 22: 3179-3190.

Auton A, McVean G. 2007. Recombination rate estimation in the presence of hotspots. Genome Res. 17: 1219-1227.

Baudry E, Depaulis F. 2003. Effect of misoriented sites on neutrality tests with outgroup. Genetics 165: 1619-1622.

Bomblies K, Madlung A. 2014. Polyploidy in the *Arabidopsis* genus. Chromosome Res. 22: 117-34.

Brochmann C, Elven R. 1992. Ecological and genetic consequences of polyploidy in arctic *Draba* (Brassicaceae). Evol Trends Plants 6: 111-124.

Cosgrove D, Bedinger P, Durachko D. 1997. Group I allergens of grass pollen as cell wall-loosening agents. Proc Natl Acad Sci USA 94: 6559-6564.

Cui L, Wall P, Leebens-Mack J, Lindsay B, Soltis D, Doyle J, Soltis P, Carlson J, Arumuganathan K, Barakat A, et al. 2006. Widespread genome duplications throughout the history of flowering plants. Genome Res. 16: 738-749.

Cutler D, Jensen J. 2010. To pool, or not to pool? Genetics 186: 41-43.

Dehal P, Boore J. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol 3: e314.

Derome N, Métayer K, Montchamp-Moreau C, Veuille M. 2004. Signature of selective sweep associated with the evolution of sex-ratio drive in *Drosophila simulans*. Genetics 166: 1357-1366.

Excoffier L, Dupanloup I, Huerta-Sánchez E, Sousa V, Foll M. 2013. Robust demographic inference from genomic and SNP data. PLoS Genet 9: e1003905.

Fay J, Wu C. 2000. Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

Fu Y. 1995. Statistical properties of segregating sites. Theor Popul Biol. 48: 172-197.

Gautier M, Gharbi K, Cezard T, Foucaud J, Kerdelhué C, Pudlo P, Cornuet J, Estoup A. 2013. The effect of RAD allele dropout on the estimation of genetic variation within and between populations. Mol Ecol. 22: 3165-3178.

Grant V. 1981. Plant speciation, 2nd edition. New York: Columbia University Press.

Henry I, Dilkes B, Young K, Watson B, Wu H, Comai L. 2005. Aneuploidy and genetic variation in the *Arabidopsis thaliana* triploid response. Genetics 170: 1979-1988.

Hill R. 1971. Selection in Autotetraploids. Theor Appl Genet. 41: 181-186.

Hohmann N, Schmickl R, Chiang T, Lucanova M, Kolar F, Marhold K, Koch M. 2014. Taming the wild: resolving the gene pools of non-model Arabidopsis lineages. BMC Evol Biol 14: e224.

Hollister J, Arnold B, Svedin E, Xue K, Dilkes B, Bomblies K. 2012. Genetic adaptation associated with genome-doubling in autotetraploid *Arabidopsis arenosa*. PLoS Genet 8: e1003093.

Hu T, Pattyn P, Bakker E, Cao J, Cheng J, Clark R, Fahlgren N, Fawcett J, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet. 43: 476-481.

Jiao Y, Wickett N, Ayyampalayam S, Chanderbali A, Landherr L, Ralph P, Tomsho L, Hu Y, Liang H, Soltis P. 2011. Ancestral polyploidy in seed plants and angiosperms. Nature 473: 97-100.

Jombart T, Ahmed I. 2011. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. Bioinformatics 27: 3070–3071.

Jørgensen M, Ehrich D, Schmickl R, Koch M, Brysting A. 2011. Interspecific and interploidal gene flow in Central European *Arabidopsis* (Brassicaceae). BMC Evol Biol. 11: 346.

Junier T, Zdobnov E. 2010. The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell. Bioinformatics 26: 1669-1670.

Kellis M, Birren B, Lander E. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature 428: 617-624.

Kingman J. 1982. On the genealogy of large populations. J Appl Probab. 19: 27-43.

Kolar F, Lucanova M, Zaveska E, Fuxova G, Mandakova T, Spaniel S, Senko D, Svitok M, Kolnik M, Gudzinskas Z, et al. 2015. Ecological segregation does not drive the intricate parapatric distribution of diploid and tetraploid cytotypes of the *Arabidopsis arenosa* group (Brassicaceae). Biol J Linn Soc. AOP 2/24/15.

Kolnik M. 2007. Arabidopsis. In: Marhold K, Martonfi P, Mereda P, Mraz P, editors. Chromosome number survey of the ferns and flowering plants of Slovakia. Bratislava: VEDA. P. 94-102.

Lunter G, Goodson M. 2011. Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 21: 936-939.

Luo J, Gao Y, Ma W, Bi X, Wang S, Wang J, Wang Y, Chai J, Du R, Wu S, et al. 2014. Tempo and mode of recurrent polyploidization in the *Carassius auratus* species complex (Cypriniformes, Cyprinidae). Heredity 112: 415-427.

Masterson J. 1994. Stomatal size in fossil plants: evidence for polyploidy in majority of angiosperms. Science 264: 421-424.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K. Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. 2010. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297-1303.

Messer P, Petrov D. 2013. Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol Evol. 28: 659-669.

Novembre J, Stephens M. 2008. Interpreting principal component analyses of spatial population genetic variation. Nat Genet. 40: 646-649.

Parisod C, Holderegger R, Brochmann C. 2010. Evolutionary consequences of autopolyploidy. New Phytol. 186: 5-17.

Peterson B, Weber J, Kay E, Fisher H, Hoekstra H. 2012. Double digest RADseq: An inexpensive method for de novo SNP discovery and genotyping in model and non-model species. PLoS ONE 7: e37135.

Pickrell J, Pritchard J. 2012. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genetics 8: e1002967.

Pritchard J, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. Genetics 155: 945-959.

Raineri E, Ferretti L, Esteve-Codina A, Nevado B, Heath S, Pérez-Enciso M. 2012. SNP calling by sequencing pooled samples. BMC Bioinformatics 13: 239.

Ramsey J, Schemske D. 1998. Pathways, mechanisms, and rates of polyploidy formation in flowering plants. Annu Rev Ecol Syst. 29: 467-501.

Ronikier M. 2011. Biogeography of high-mountain plants in the Carpathians: An emerging phylogeographical perspective. Taxon 60: 373-389.

Schmickl R, Paule J, Klein J, Marhold K, Koch M. 2012. The evolutionary history of the Arabidopsis *arenosa* complex: Diverse tetraploids mask the Western Carpathian center of species and genetic diversity. PLoS ONE 7: e42691.

Segraves K, Thompson J, Soltis P, Soltis D. 1999. Multiple origins of polyploidy and the geographic structure of *Heuchera grossulariifolia*. Mol Ecol. 8: 253-262.

Soltis D, Buggs R, Doyle J, Soltis. 2010. What we still don't know about polyploidy. Taxon 59: 1387-1403.

Soltis D, Soltis P, Ness B. 1989. Chloroplast-DNA variation and multiple origins of autopolyploidy in *Heuchera micrantha*. Evolution 43: 650–656.

Ståhlberg D. 2009. Habitat differentiation, hybridization and gene flow patterns in mixed populations of diploid and autotetraploid *Dactylorhiza maculata* s.l. (Orchidaceae). Evol Ecol. 23; 295-328.

Stebbins G. 1950. Variation and evolution in plants. New York: Columbia University Press.

The Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408: 796–815.

Thórsson T, Salmela E, Anamthawat-Jónsson K. 2001. Morphological, cytogenetic, and molecular evidence for introgressive hybridization in birch. J Hered. 92: 404-8.

Van Dijk P, Bakx-Schotman T. 1997. Chloroplast DNA phylogeography and cytotype geography in autopolyploid *Plantago media*. Mol Ecol. 6: 345-352.

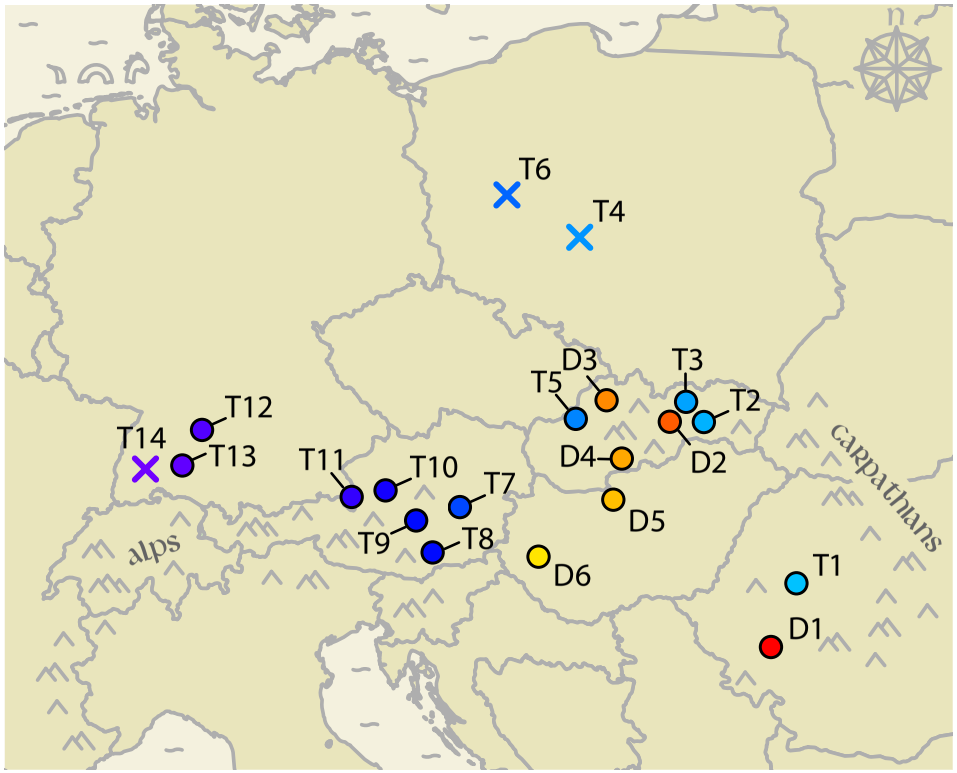
Wood T, Takebayashi N, Barker M, Mayrose I, Greenspoon P, Rieseberg L. 2009. The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci USA 106: 13875–13879.

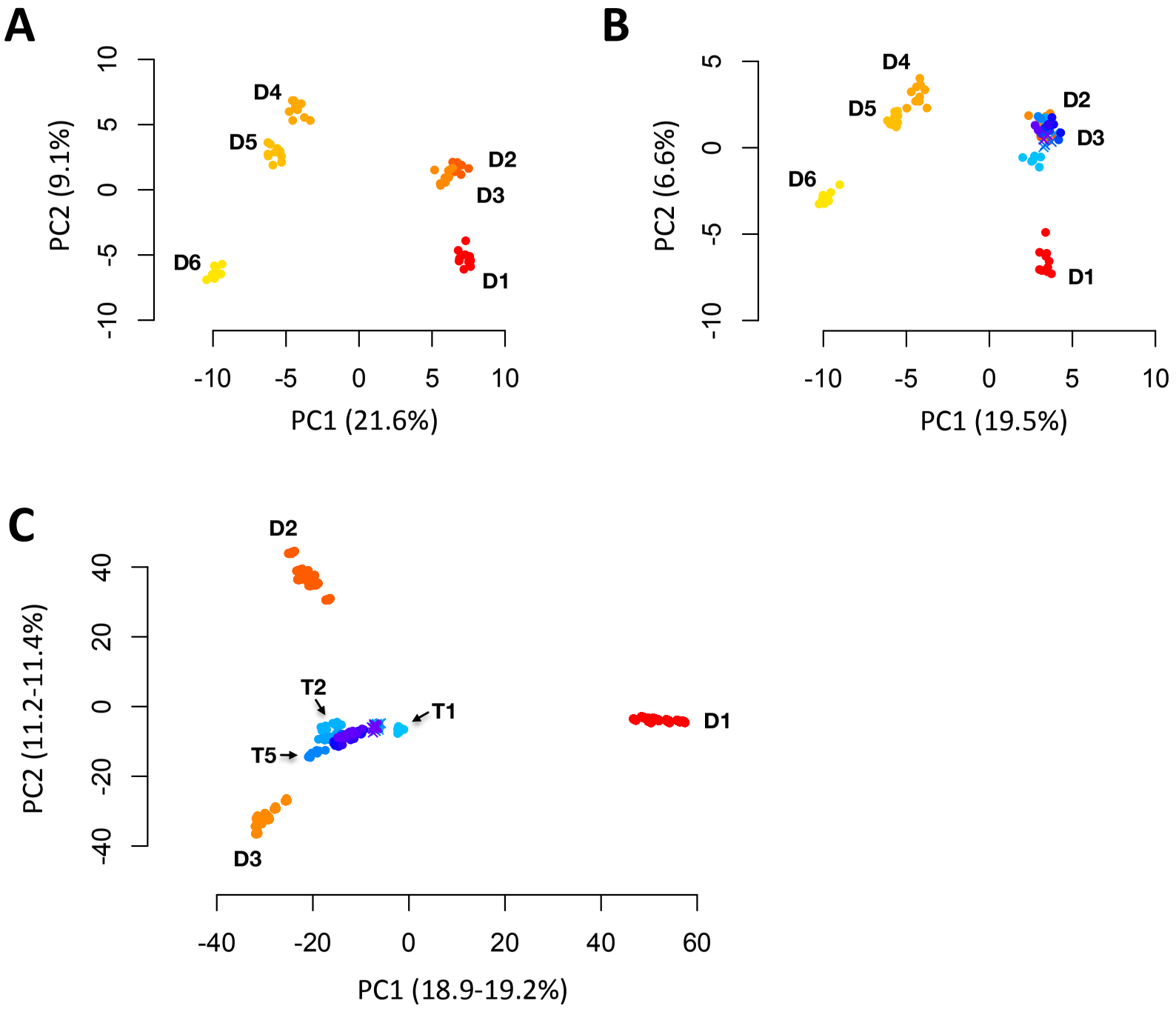
Wright K, Arnold B, Xue K, Surinova M, O'Connell J, Bomblies K. 2015. Habitat and cytotype associated selection on meiosis proteins in Arabidopsis arenosa. Mol Biol Evol. 32: 944-955.

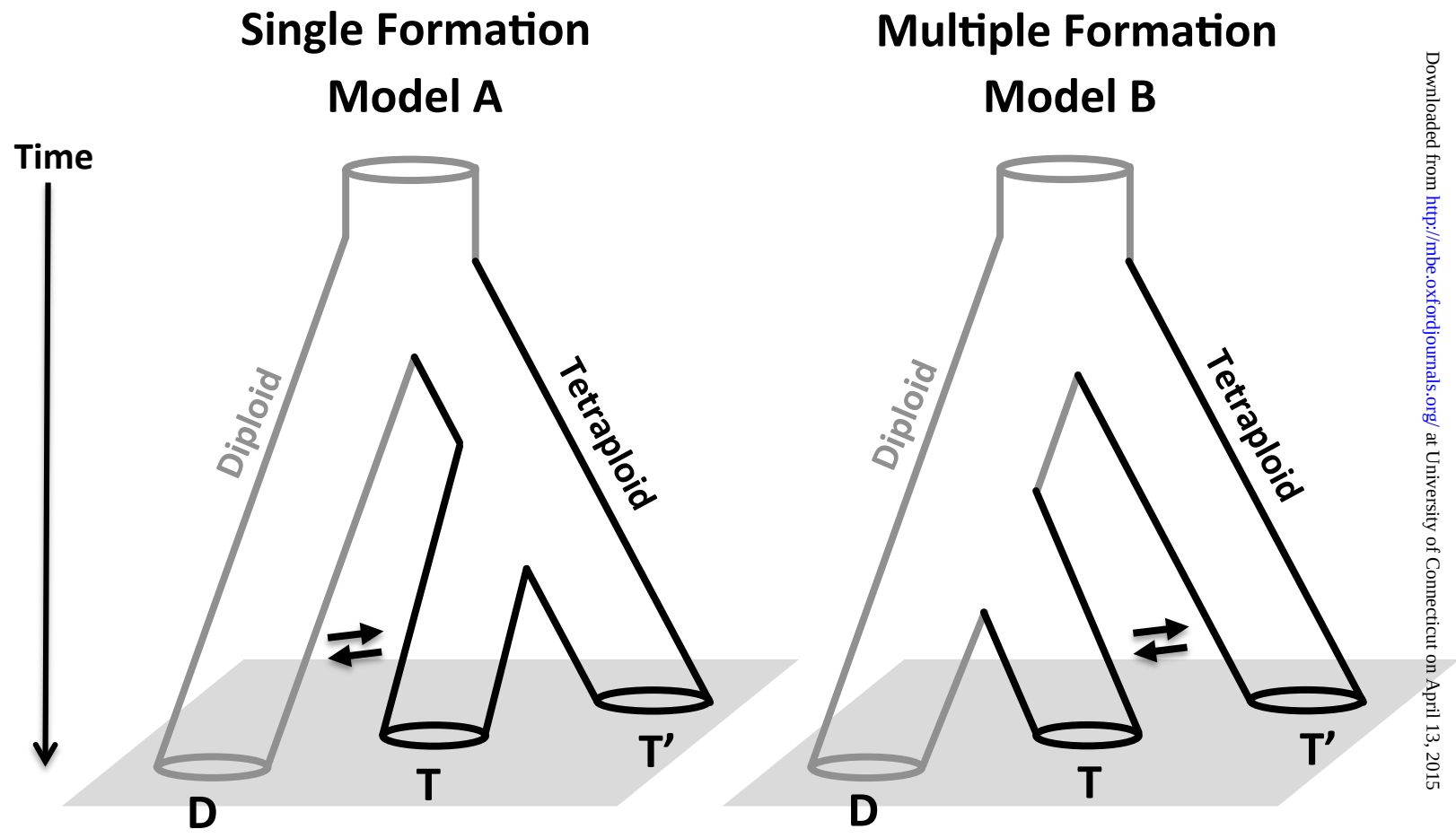
Yamane K, Yasui Y, Ohnishi O. 2003. Intraspecific cpDNA variations of diploid and tetraploid perennial buckwheat, *Fagopyrum cymosum* (Polygonaceae). Am J Bot. 90: 339-346.

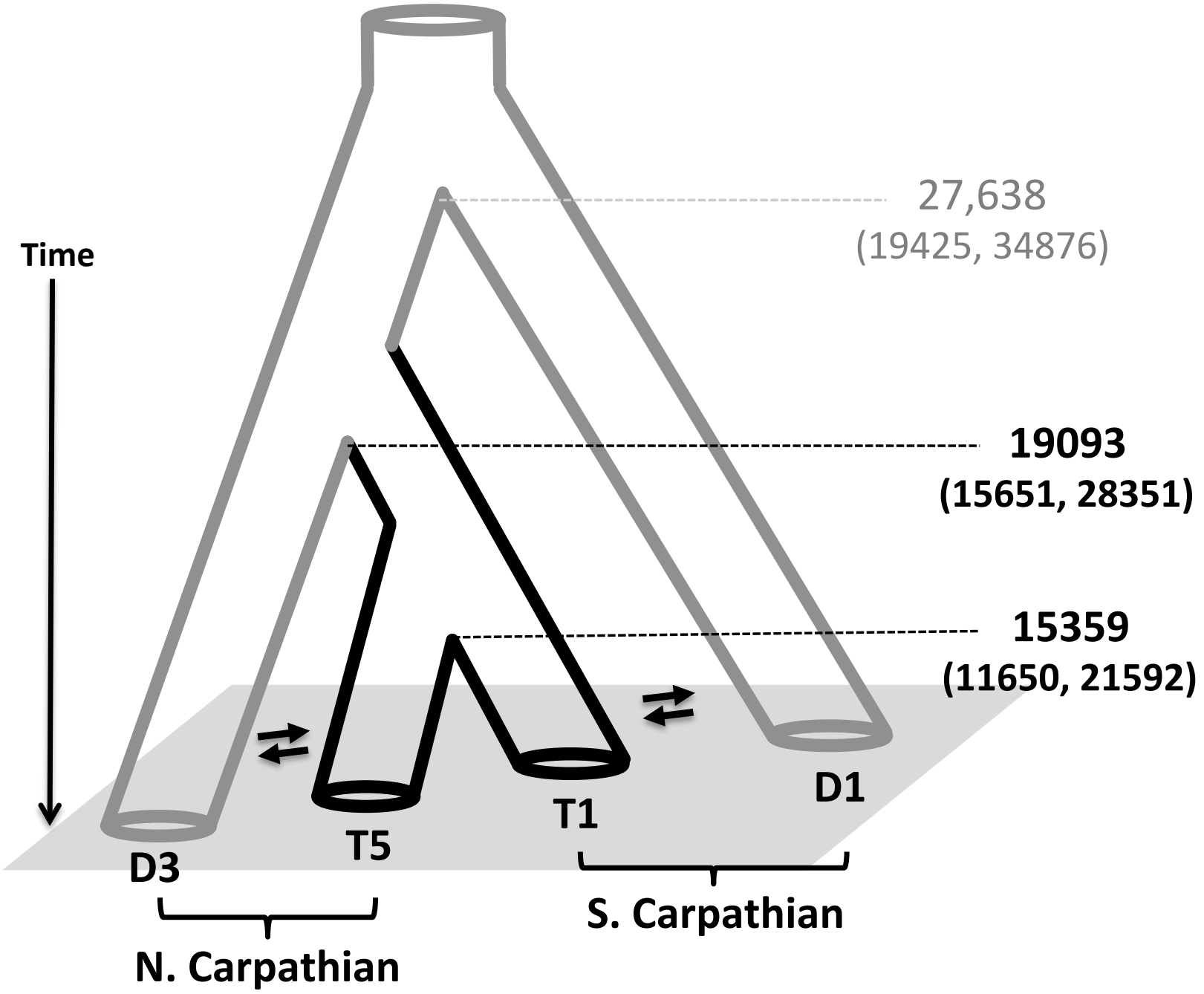
Yang W, Glover B, Rao G, Yang J. 2006. Molecular evidence for multiple polyploidization and lineage recombination in the *Chrysanthemum indicum* polyploid complex (Asteraceae). New Phytol. 171: 875-886.

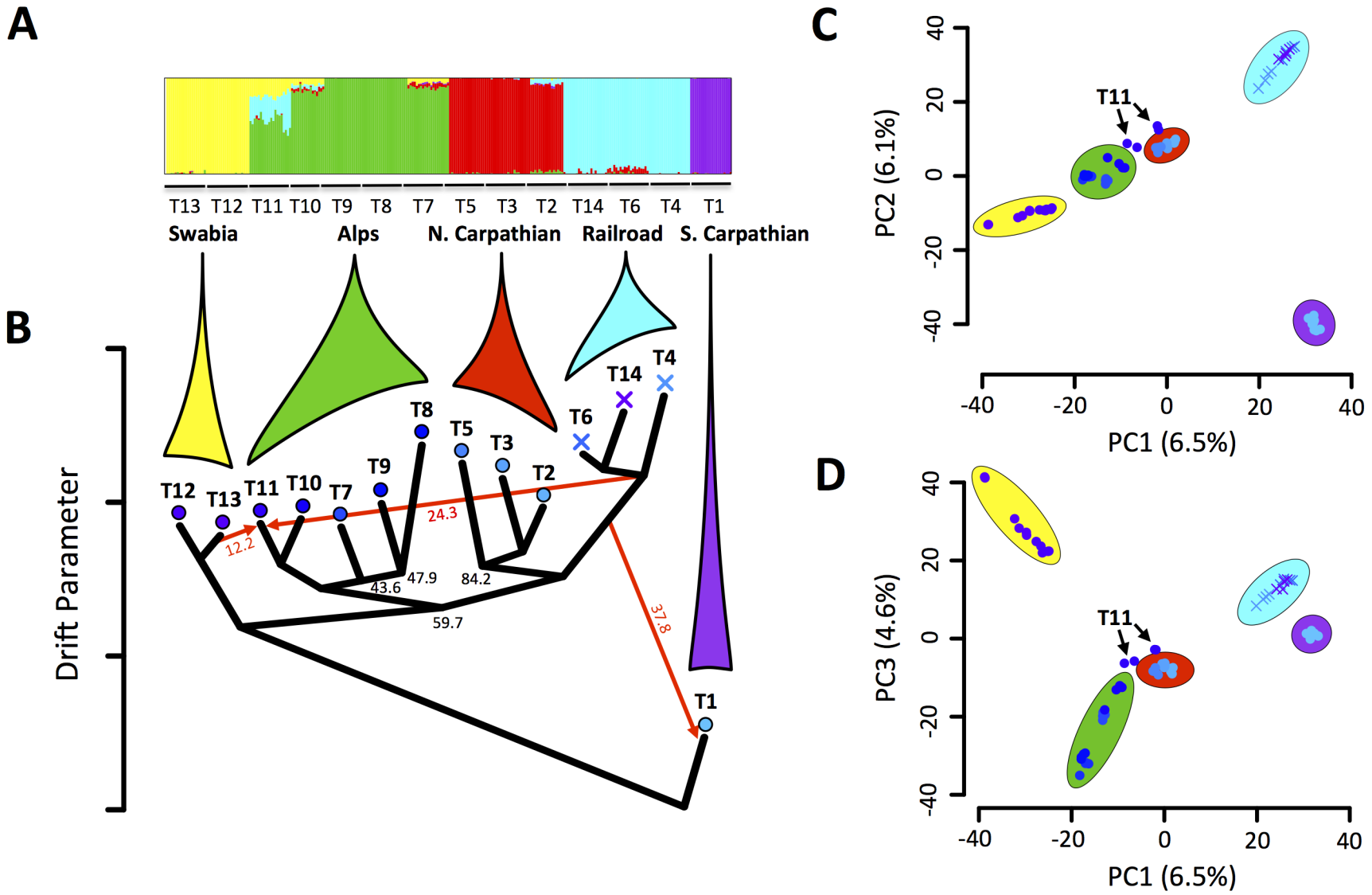
Yant L, Hollister J, Wright K, Arnold B, Higgins J, Franklin C, Bomblies K. 2013. Meiotic adaptation to genome duplication in *Arabidopsis arenosa*. Curr Biol 23: 2151-2156.

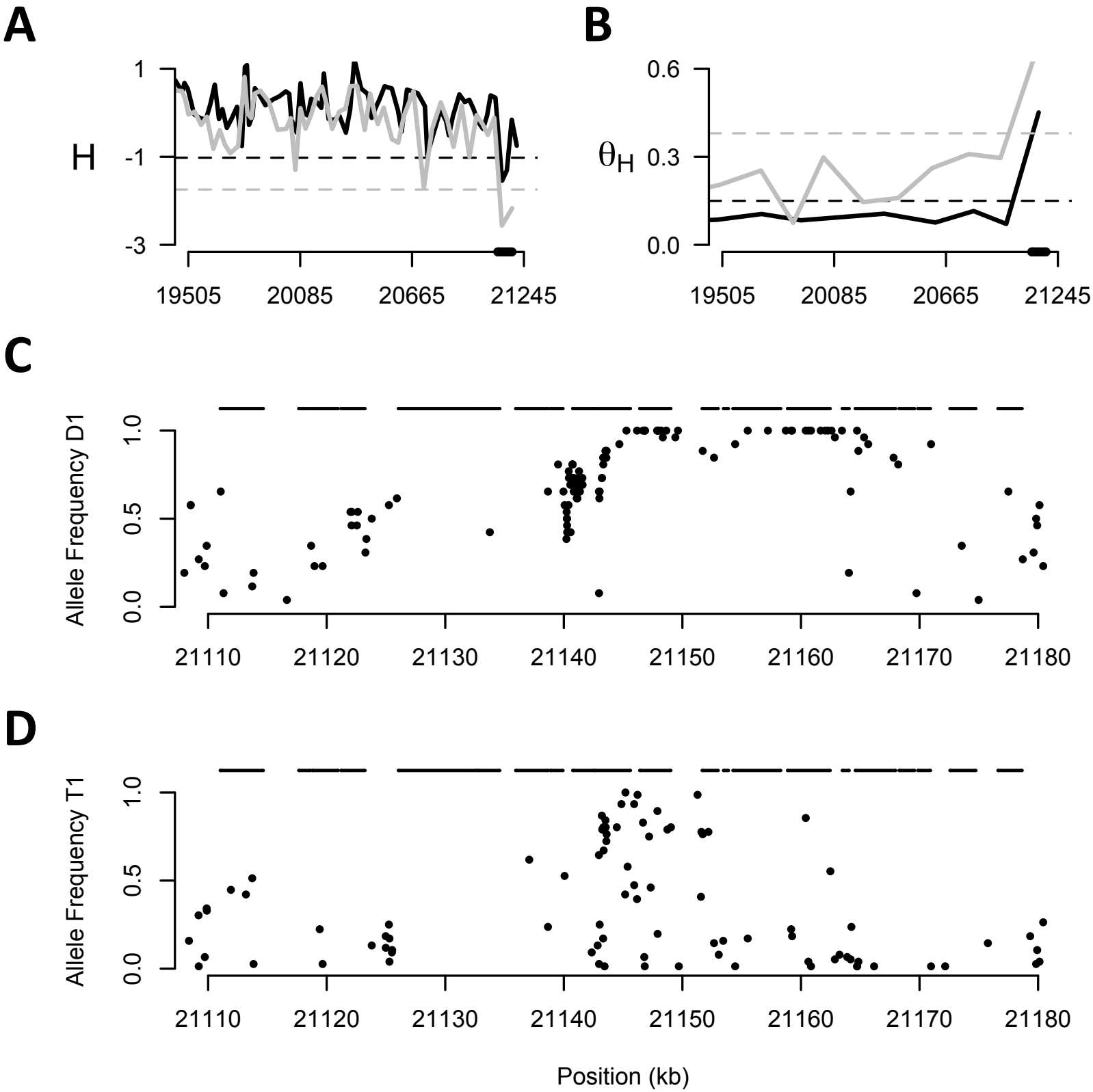







Time
27,638
(19425, 34876)
19093
(15651, 28351)
15359
(11650, 21592)
D3
T5
T1
D1
N. Carpathian
S. Carpathian


A
T13 T12 T11 T10 T9 T8 T7 T5 T3 T2 T14 T6 T4 T1
Swabia
Alps
N. Carpathian
Railroad
S. Carpathian
B
Drift Parameter
T12
T13
T11
T10
T7
T9
T8
T5
T3
T2
T14
T6
T4
T1
12.2
24.3
43.6
47.9
84.2
59.7
37.8
C
PC2 (6.1%)
PC1 (6.5%)
T11
40
20
0
-20
-40
D
PC3 (4.6%)
PC1 (6.5%)
T11


A
H
1
-1
-3
19505
20085
20665
21245
B
θH
0.6
0.3
0.0
19505
20085
20665
21245
C
Allele Frequency D1
0.0
0.5
1.0
21110
21120
21130
21140
21150
21160
21170
21180
D
Allele Frequency T1
0.0
0.5
1.0
21110
21120
21130
21140
21150
21160
21170
21180
Position (kb)